



TC2004B.200 Análisis de Ciencia de Datos

Profra. Ma. Angelina Alarcón Romero

Profra. María de los Angeles Constantino González

**“Optimización en la
Selección de Personal Destacado”**

E-Detecter

Equipo 007

Andres Saldaña Rodriguez A01721193

Karen González Ugalde A01411597

Fernando Gonzalez Rosas A01253694

José Eugenio Morales Ortiz A01734612

Daniel Sánchez Villarreal A01197699

06 / Abril /2022

Nombre	Rol
José Eugenio Morales Ortiz	Científico de Datos/ PM
Andres Saldaña Rodriguez	Científico de Datos/ Ing. de Datos
Karen González Ugalde	Científico de Datos / Ing. de Datos
Fernando González Rosas	Científico de Datos / Ing. de Datos
Daniel Sánchez Villarreal	Científico de Datos / Ing. de Datos

Comprensión del negocio

Objetivos

Objetivos a intentar alcanzar:

- Poder visualizar de manera sencilla y eficaz los resultados de aptitud de los aplicantes
- Disminuir el tiempo que se toma encontrar aplicantes destacados y resaltarlos
- Hacer más eficiente (menos tardado y tedioso) el ingreso de nuevos aplicantes a la base de datos
- Usar la “ciencia de datos” para encontrar o predecir a los aplicantes que tendrán el mejor desempeño

Evaluación de la situación actual

Actualmente la empresa de Ternium realiza su proceso de selección de nuevos aplicantes de manera manual y mecánica, dicha forma consume bastante tiempo y con el paso del tiempo el número de aplicantes aumenta, lo cual hace este proceso aún más tedioso de lo que ya es. Bajo estas tareas mecánicas también se tiene la de encontrar manualmente a posible personal destacado, que al igual que con el punto anterior, consume mucho tiempo.

Entonces lo que se busca es poder disminuir el tiempo de selección e identificación de personal destacado, además de poder asegurar de mejor manera que dicho personal sea sobresaliente.

Solución propuesta

Nuestra propuesta es diseñar una herramienta para poder optimizar la selección del personal destacado y minimizar el tiempo de identificación de este personal de la empresa Ternium. Esto lo planeamos resolver mediante el uso de la ciencia de datos, utilizando herramientas como python y excel para crear una nueva base de datos limpia en la que sea más fácil visualizar la información, ya sea en una nueva página web interactiva en la que solo tengan acceso los directivos de Ternium o directamente desde un archivo.

Hipótesis

Lo que se busca tener como resultado final es el poder desarrollar un nuevo método para el área de recursos humanos, que les permita poder realizar el proceso de selección de personal nuevo destacado y altamente calificado de una manera más automatizada y sencilla de analizar e interpretar.

Esto se planea lograr utilizando diferentes conceptos y métodos conocidos en la ciencia de datos, tales como el análisis descriptivo, análisis predictivo, correlación de variables, arquitectura de bases de datos, Business Intelligence; los cuales consideramos que nos pueden ser de utilidad para la principal meta que se busca resolver en esta problemática, la cual es la facilitación en la toma de decisiones en la selección de nuevo personal.

Justificación

El uso de ciencia de datos y análisis predictivo no es algo nuevo en este tipo de campos, varias propuestas e investigaciones han sido hechas en intentos de incorporar estos nuevos conceptos en diferentes áreas.

Por ejemplo se encuentran estudios ya hechos de análisis predictivo en el departamento de recursos humanos en el que se buscaba generar un modelo predictivo para poder deducir si algún empleado abandonaría la empresa de manera espontánea y en base a esto tomar acciones para evitar dicho resultado. [1]

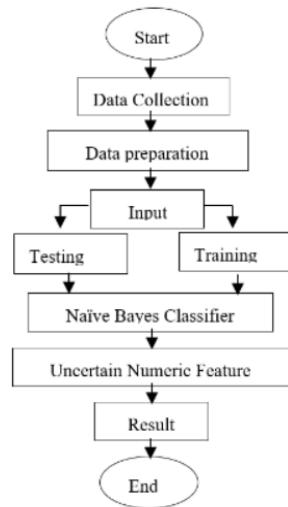


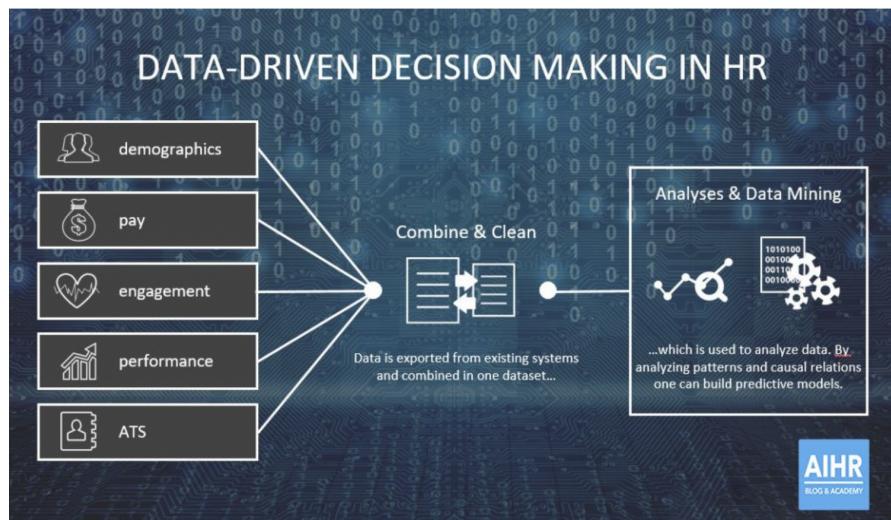
Fig. 1. Research Flowchart

Este modelo predictivo se basó en Naïve Bayes, mejor conocido como por el Teorema de independencia de Bayes y lo que buscó fue predecir si un empleado renunciaría o no, basado en sus características, pero claramente esto no fue lo único que se hizo. Desarrolló la solución del problema utilizando clasificación de factores numéricos donde la precisión de la prueba depende de la efectividad del entrenamiento del método. En dicho estudio también se muestra la tabla de flujo realizada para la resolución del problema. [1]

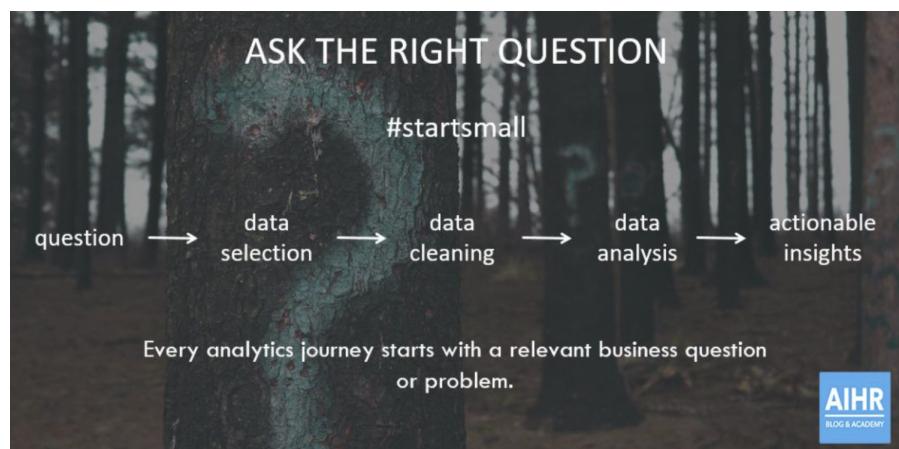
Este es un buen ejemplo de cómo estos conocimientos ya están siendo utilizados dentro del mundo laboral ya que a lo largo del documento de referencia se encuentran conceptos, como análisis descriptivo, predictivo y prescriptivo; se tiene la recolección y preparación de los datos para luego proceder a su experimentación. Si quisieran leer el resto del documento pueden checarlo en las referencias encontradas al final del documento.

En los años recientes, se ha estado mencionando el concepto de “HR Analytics”, el cual es la idea de tener una orientación más dirigida hacia los datos dentro del área de recursos humanos. De esta manera se involucrarían menos los sentimientos y serían los datos, los encargados de tomar las decisiones difíciles. [2]

Sin embargo, lo más importante para poder llevar a cabo esta transformación del área de Recursos Humanos, es muy necesario tener medidas cuantificables de éxito, o de desempeño. Es decir, que se pueda expresar con números, si un empleado tiene un buen desempeño de la empresa. Aunque suena como un acto difícil, hay maneras de conseguirlo. [2]



Esta innovación dentro del área, trae consigo posibilidades muy emocionantes. Como la capacidad de predecir cuáles empleados van a ser los mejores dentro de la compañía, ver cuáles empleados son los más probables a renunciar o fracasar dentro de la compañía. Muchos avances que cambiarían el proceso de reclutamiento que lleva décadas siendo el mismo. [2]



AIHR ofrece distintos cursos para capacitarse dentro de esta área de “HR Analytics”. Aunque no vayamos a tomar los cursos, es muy interesante ver cómo no es complicado integrar la cuenca de datos con el área de recursos humanos. Aunque antes se pudiera creer que esto sería muy complicado, y que no había espacio para la ciencia de datos en algo tan viejo y anticuado como lo son los recursos humanos, ahora hasta hay cursos para que cualquiera pueda aprender a hacerlo.

Ante la demanda acelerada de nuevas habilidades, se requiere para responder de usar la tecnología para la recolección de datos, su protección y el análisis adecuado, con el fin de encontrar el compromiso, ética, capacitación y habilidades e iniciativas más amplias del talento humano reclutado. [3]

No en vano, un reciente estudio realizado por UBITS “El futuro del trabajo en Latinoamérica 2020” evidenció que el 72.1% de los gerentes de RR.HH están implementando nuevas tecnologías para agilizar los procesos de selección y reclutamiento. Y su uso con el fin de optimizar tales procesos se aceleró un 67.3%. [3]

La implementación de estos programas tecnológicos le permiten al reclutador gestionar todas las etapas del proceso de forma efectiva y ágil, pero también mantener una comunicación constante de tal forma que los procesos sean confiables. Es así como las plataformas avanzadas y los análisis basados en la captura de datos y estadísticas están enfocados en encontrar perfiles aptos dentro de la escasez de talento adecuado que está presentando el mercado laboral actual. Por lo que se implementan pruebas especializadas que brindan información precisa de forma ágil para dar respuestas rápidas sobre los candidatos que se están evaluando y captarlos rápidamente, antes que la competencia. [3]

El futuro cercano parece regido por habilidades que antes no tenían la misma importancia crítica y estratégica en apariencia, como también por modelos laborales y operativos, de allí la importancia de que las empresas puedan liderar estrategias de reclutamiento que integren todos procesos clave para la efectividad y contención de colaboradores adaptados y productivos en estos escenarios, en pro de una recuperación económica rápida y sostenida, en la que la capacitación, trabajo remoto, bienestar y

comunicación resultan ser esenciales desde el momento cero de verdad en la selección y reclutamiento. [3]

La analítica de recursos humanos (HR Analytics) es una práctica que a lo largo de los años se ha estado introduciendo lentamente en instituciones y empresas con la necesidad de tener herramientas predictivas basadas en metodologías no sin sesgos humanos para la toma de decisiones a corto, mediano y largo plazo.

Este análisis puede ser descrito bajo el siguiente proceso: identificación del problema, realización de una investigación acerca del modelo específico de recursos humanos para la problemática, manejo de los datos, análisis datos, interpretación y comunicación de los resultados, y por último, un plan de acción consecuente con este proceso [4]. El proceso descrito es notablemente sencillo, lo cual, nos hace preguntarnos sobre porque, aun en la actualidad, esta metodología no es utilizada de manera generalizada, y, posiblemente es debido a la falta de reconocimiento que se tiene a los resultados que este puede proporcionar. [5]

Estudios han mostrado que las compañías u organizaciones que hacen uso de ciencia de datos, análisis de datos, y herramientas y plataformas tecnológicas para la utilización de Big Data en su gestión, gastan aproximadamente un 20% menos en la gestión de recursos humanos por cada empleado contratado de este modo [6]. En Irlanda, de un total de 81 líderes de recursos humanos en Irlanda, se halló que el análisis de datos era la prioridad número uno, y hay una tendencia de que los profesionales que cuentan con experiencia en análisis tienen una ventaja sobre quienes no la tengan, y el análisis y la ciencia de datos son cada vez más importantes y más solicitados y utilizados en el mundo de recursos humanos.

El hacer uso de herramientas tecnológicas basadas en datos y análisis de datos ofrece la posibilidad y beneficio de tener una mayor comprensión, representación y monitoreo de información y medidas que pueden ser de ayuda para poder responder de

manera clara y objetiva a diversas cuestiones de mucha importancia, tanto para recursos humanos como para la compañía u organización en sí, así como para quienes se dedican a realizar esto. Por ejemplo se puede conocer y comunicar con más claridad y certeza ¿cuántas horas de capacitación se invierten en promedio para cada empleado, o para cada equipo contratado?, ¿cuánto tarda en promedio la incorporación de un nuevo empleado?, etc. Además de que esto puede ayudar a que el proceso se vuelva menos tardado e incluso puede que para llevar a cabo este proceso ya no se ocupe tanto personal como el que se podría ocupar si no se hace uso de estas herramientas, como sucedía antes hace algunos años, por lo que además de hacer que el proceso se vuelva más eficiente, también podría verse reflejada una baja o reducción de costos o gastos al llevar a cabo el proceso.

Los tableros de gestión combinados con el uso de analítica ofrecen a los especialistas en recursos humanos la posibilidad de una mayor facilidad y eficiencia para mostrar y evaluar desempeño o rendimiento, monitorear y predecir, y tomar decisiones efectivas que brindan valor a los empleados y a la compañía. Con esto se puede reducir u optimizar costos de recursos humanos, atraer a los empleados con mejor talento, y ayudar a conservar a los empleados más valiosos.

La ciencia de datos les sirve a las empresas y compañías para determinar patrones, detectar problemas, e identificar posibles riesgos. Al utilizarla, es posible analizar una gran cantidad de datos, relacionarlos, y mostrar tendencias, patrones y lógicas que a simple vista pueden aparentar no relacionarse, y con esto se pueden obtener conclusiones de valor sobre los empleados y varios aspectos con respecto a ellos, como su satisfacción en la empresa, su desempeño, rendimiento y muchas más. Para esto se pueden utilizar modelos predictivos, como en el caso de un modelo de renuncia, el cual indica en qué porcentaje un empleado es propenso a renunciar a la empresa o compañía a corto plazo; así como se puede usar en muchas situaciones más.

Mercado Potencial

Lo más emocionante de este proyecto, aparte del proyecto en sí, son las posibilidades. Ya que teóricamente puede ser usado por cualquier compañía. Es un esquema que se puede adaptar muy fácilmente, aunque ahorita va a ser específico para Ternium, en sí la estructura, y la idea detrás del proyecto es muy fácil de adaptar a cualquier otra compañía. Siempre y cuando tengan una base de datos amplia de sus empleados, así como distintas pruebas o métricas para poder evaluarlos.

Suponiendo que seamos exitosos en la entrega final del prototipo, sería muy fácil venderlo a otras compañías. Así que nuestro mercado potencial es muy amplio. Claro que no podemos tomar en cuenta empresas pequeñas, porque estas no tienen una gran cantidad de empleados. Pero cualquiera de las empresas moderadamente grandes sería un balnco perfecto para nosotros. Incluso si no tienen ningún método de recopilación de datos, o llegan a tener bases de datos muy simples, estas se pueden modificar como parte de la propuesta para adaptarlas a llevar más métricas de empleados. En realidad lo único que necesitamos es que tengan una gran cantidad de empleados, es el único requerimiento. Por esto mismo nuestro mercado potencial es inmenso, porque el requerimiento no es complicado de cumplir.

Para darnos una idea de este mercado, el 0.8% de las empresas PYME de méxico son medianas, lo que significa que tiene entre 50-250 personas empleados, estas serían nuestro blanco principal mínimo. [7]

Considerando que hay un total aproximado de 4.1 millones de empresas PYME's en México [7], hay un aproximado de 32,800 empresas potenciales, que clasifican como PYMES Medianas, a las que podríamos ofrecer nuestro proyecto.

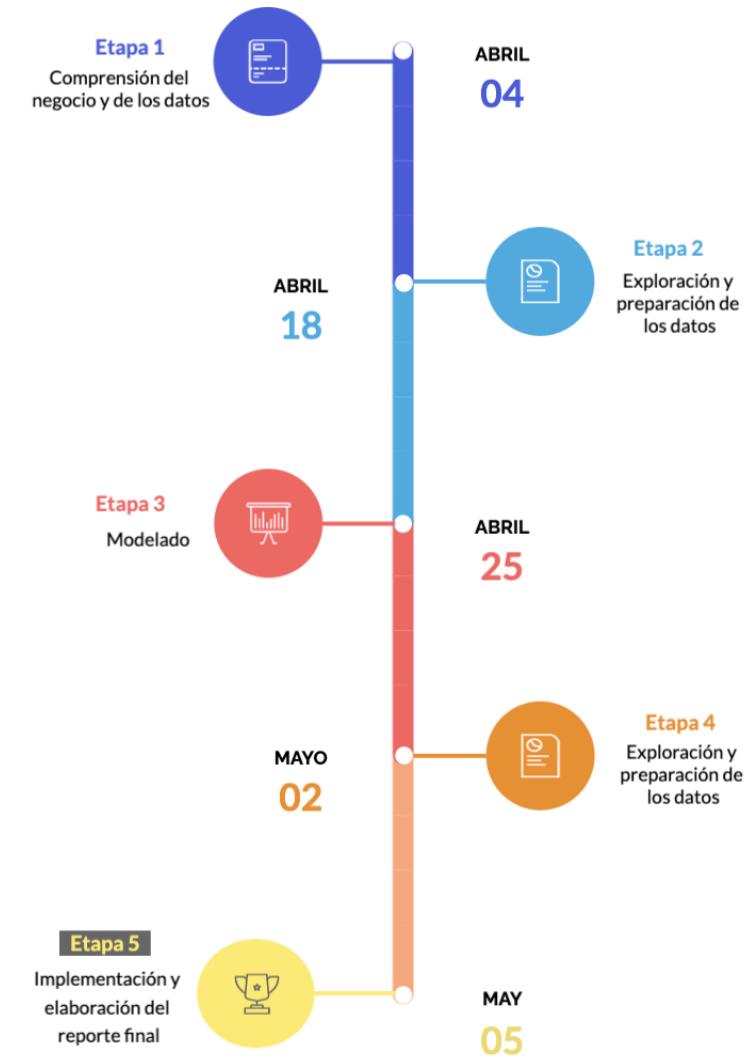
Identificación de clientes/consumidor y usuarios

Una buena ventaja que tiene el desarrollo de esta solución es que si bien en este momento se encuentra concentrada para resolver la problemática de recursos humanos de Ternium, debido a la estructura en la cual estaría hecha la solución nos da la ventaja de que este mismo modelo podría ser utilizado de manera general para cualquier

organización que quiera mejorar sus cuestiones de recursos humanos y reclutamiento de personal sobresaliente.

La única diferencia sería que como no todas las empresas evalúan a los aplicantes de la misma manera lo que se tendría que hacer es modificar modelo y ajustarlo acorde a las necesidades, lineamientos, reglas y criterios de evaluación de la organización que quiera utilizar esta solución. De esta manera no nos limitamos solamente a Ternium como cliente, sino que ahora existe la posibilidad de poder expandir nuestro a nuestros posibles clientes a las empresas que busquen una manera más automatizada para la selección de un personal excepcional.

Plan de actividades del proyecto



Equipo 7	
Andres Saldaña Rodriguez	A01721193
Karen González Ugalde	A01411597
Fernando Gonzalez Rosas	A01253694
José Eugenio Morales Ortiz	A01734612
Daniel Sánchez Villarreal	A01197699

Comprensión de los datos

Descripción del set de datos

Podemos empezar viendo la base de datos actualizada por nuestro socio formador Ternium que utilizan para la selección de su nuevo personal y practicantes. La base de datos presentó más de 5000 filas queriendo decir que hubo más de 5000 aplicantes para puestos de trabajo y prácticas en Ternium, a su vez se tienen 51 columnas en la base de datos que esto lo que nos dice es que hay 51 atributos que se registran acerca de los participantes, la variedad de ellos es bastante y se tienen varias cosas como la nacionalidad, ID, correo electrónico, semestres cursados, nivel de inglés, tipo de evento en el que se enteraron, actividades grupales, las prueba pym metrics, entre muchas otras más.

Después de considerar las variables que se tenían a la mano y dialogar acerca de cuales no resultaban de utilidad para la solución que queríamos proporcionar, terminamos con un total de 20 variables las cuales mostramos a continuación.

- **País:** Nacionalidad
- **ID Candidato:** Personal
- **Género:** Hombre o Mujer
- **Carrera Gestional:** Área de carrera decidida por la empresa
- **Avance:** Cuantos semestres llevan de carrera
- **Semestres Totales:** Cuantos semestres dura su carrera
- **Postulados Sí/No:** Sí se postuló para la prueba o no
- **Evaluados Sí/No:** Sí realizó las pruebas o no
- **Altamente Recomendado:** Sí se considera altamente recomendado
- **Operaciones-Calidad:** Prueba pym metrics
- **MTTO-DIMA:** Prueba pym metrics
- **Comercial-Planeamiento:** Prueba pym metrics
- **DIGI-SC:** Prueba pym metrics
- **Resto-Soft:** Prueba de soft-skills
- **Apto/No Apto:** Si se considera apto o no

- **Destacado Pym:** Sí se considera destacado o no
- **Ingles:** Nivel de inglés con rango desde A1-C2
- **Apto:** Sí se considera apto en inglés
- **Destacado:** Sí se considera destacado en inglés
- **Ingresados Si/No:** Si el aplicante ingreso o no

Cabe mencionar que esto es tan solo nuestra elección inicial de nuestras variables de trabajo, cabe la posibilidad de que mientras siga el desarrollo del proyecto consideremos el eliminar más variables si es que lo vemos conveniente.

Es importante también hablar de las reglas que maneja Ternium en cuanto a los resultados de las pruebas pym metrics que son las de mayor importancia en el proceso de selección.

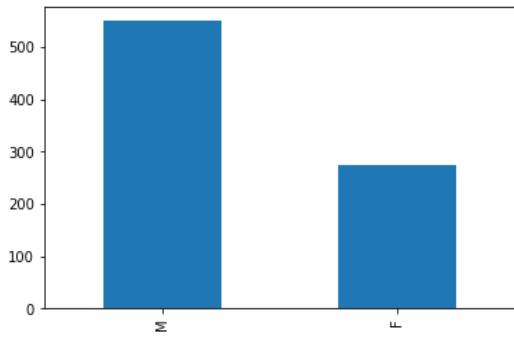
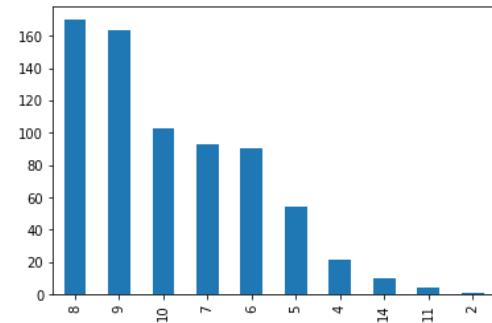
Las pruebas Pymetrics sirven para determinar la viabilidad del aplicante en las áreas, y la última prueba es usada para determinar las soft-skills que se pueden definir como una mezcla de competencias sociales, atributos personales, cualidades y atributos que definen el desempeño de una persona en su entorno. Cada una de estas pruebas se puede calificar de tres formas: “Do Not Recommend”, “Recommend” “Highly Recommend”. La forma de criterio de evaluación que ellos tienen es la siguiente.

- Si en alguna de las primeras 4 pruebas se obtiene “Highly Recommend” entonces se le clasifica como “Destacado”.
- Si únicamente sale como “Recommend” se le clasifica como “Apto”
- Si sale como “Highly Recommend” en las pruebas de Soft-skills se le clasifica como “Apto”
- Si en todas las pruebas sale como “Do Not Recommend” se le clasifica como “No Apto”

Luego en la sección de inglés como bien se mencionó se tiene desde el nivel A1 hasta el C2 pero también se encuentran valores como “False beginner” y “Late hangup”, estos valores serán tratados de manera correspondiente cuando se tenga que hacer la modificación de los datos.

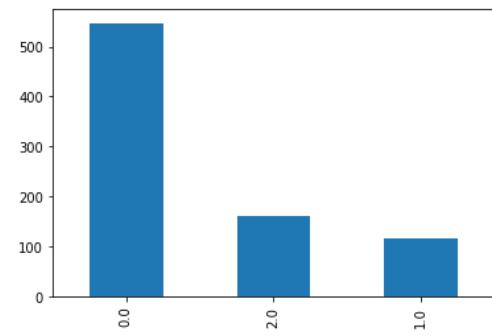
De momento esto sería todo en la descripción de los datos, a continuación pasaremos a mostrar las modificaciones hechas sobre la base para poder facilitar nuestra interpretación de los datos.

Con esto realizado podemos hacer una pequeña exploración de los datos para poder darnos un contexto de la situación, por ejemplo podemos ver que tipo de estudiantes son los que más solicitan a puestos en Ternium, esto lo podemos ver con un gráfico de barras de la variable Avance.

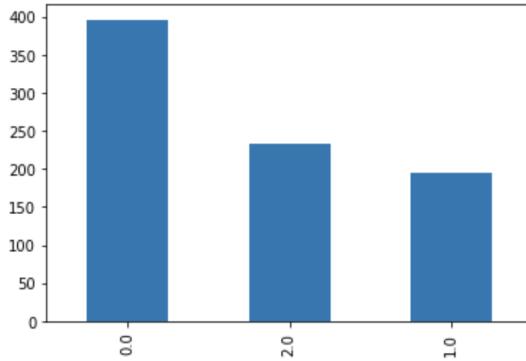


Entonces podemos observar que los alumnos que más suelen aplicar son quienes se encuentran en octavo semestre de la carrera lo cual hace sentido para empezar a hacer prácticas profesionales. Otro gráfico básico que podemos ver es cuál es la diferencia entre aplicantes hombres y mujeres.

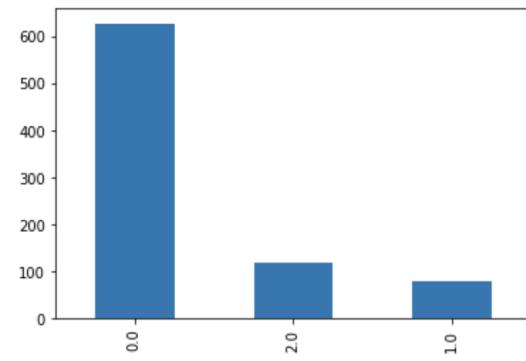
Entonces lo que concluimos rápidamente es que más hombres son los que aplican que las mujeres, la razón de por que se de esto requeriría más investigación. Podemos mostrar como último los resultados de las pruebas pym metrics las cuales son de los factores decisivos para escoger personal.



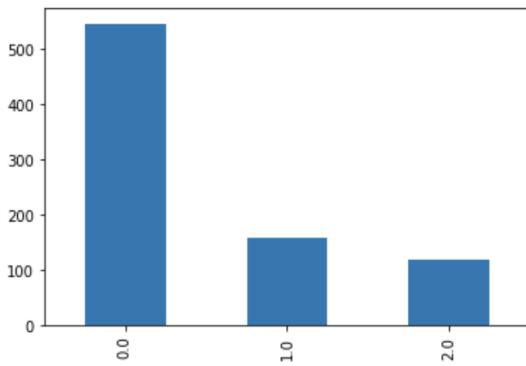
Prueba 'Operaciones-Calidad'



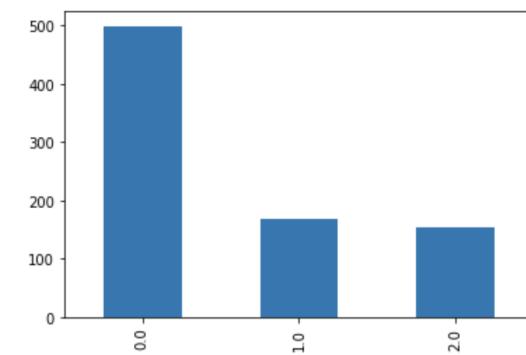
Prueba 'MTTO-DIMA'



Prueba 'Comercial-Planeamiento'



Prueba 'DIGI-SC'



Prueba 'Rest-Soft'

Podemos ver que en todos los casos el resultado dominante es 0 lo cual significa no recomendado, la variación que hay entre estos casos son la ventaja que este resultado lleva sobre otros, por ejemplo en la prueba MTTO-DIMA la ventaja que lleva no es tan grande e incluso mucha gente salió como destacado en esa categoría, pero en general se puede considerar alarmante la gran cantidad de personas que califican como no recomendado en los resultados de las pruebas.

Como último podemos mostrar brevemente los resultados estadísticos de nuestra variable de inglés.

```

count      124.000000
mean       2.911290
std        1.776139
min        0.000000
25%        2.000000
50%        3.000000
75%        4.000000
max        6.000000
Name: Ingles, dtype: float64

```

Entonces como una interpretación rápida podemos ver que el promedio de nivel de inglés que los aplicantes tienen es de 2.9, esto se traduce a que tienen un nivel entre A2 y B1, siendo más cercano al B1 (estos valores numéricos son explicados en el apartado de preparación de los datos), luego en los percentiles tenemos en el 75% al 4, esto nos quiere decir que el 75% de nuestros aplicantes tiene como máximo B2 de nivel, esto se puede considerar como algo bueno ya que B2 suele ser considerado como el mínimo indispensable. Algo importante a mencionar aquí es que estos datos son tomados de las 124 personas que tomaron las examinaciones de inglés, esto como tal representa tan solo una fracción de nuestra base de datos pero nos puede ayudar a darnos una idea general como se encuentran los aspirantes mexicanos en cuanto a niveles de inglés.

Con eso concluiríamos de momento nuestra exploración de los datos, esto se hizo meramente para poder darnos una contextualización de la situación de Ternium y en esperanzas que una representación gráfica los haya ayudado a entender de una manera más clara cómo se encuentra su empresa.

Preparación de los datos

Limpieza y Transformación de los datos

Entonces podemos empezar ya a modificar las variables que estaremos utilizando en orden de mejorar nuestra capacidad de manejar nuestra información, también en caso de que en la descripción de los datos se haya omitido algún tipo de información que fuese importante explicar puede que esta sección aclare dichas dudas.

Empecemos entonces por la variable de Nacionalidad, a esta variable planeábamos originalmente no modificarla pero dado que se nos pide trabajar momentáneamente con sólo personas de nacionalidad mexicana lo que se hizo fue eliminar las filas de esa columna en las que la nacionalidad fuera distinta, también como ya únicamente estamos trabajando con personas de nacionalidad mexicana llegamos a la conclusión de que no hay necesidad de mantener la columna por la cual decidimos eliminarla.

Después pasamos a trabajar con la variable del ID del aplicante, se dio el caso en el que varios de los aplicantes no tenían un identificador pero si realizaron las pruebas y borrar a las filas sin ID nos quitaría mucha información, entonces a la solución a la que llegamos fue que se reemplazarían los ID existentes y se les asignaría uno nuevo y a su vez darles un ID a las personas que previamente no contaban con uno.

A la variable de género y carrera gestional llegamos a la conclusión de que no requieren modificación y se dejaron como estaban. Después tenemos a las variables de Avance y semestres totales donde son tratados como clase objeto lo cual nos puede presentar problemas a la hora de querer analizar los datos, así que decidimos cambiarlo a un tipo de valor numérico para poder análisis estadísticos sobre ellos. Posteriormente como hay valores que se encuentran en palabras tales como "Egresado" y "12 o más" decidimos darle a estas variables un valor numérico categórico, a "Egresado" = 14 y "12 o más" =13, escogimos estos números para no crear outliers que nos generen más problemas en el futuro.

Para las variables de postulados y evaluados tuvimos que reemplazar los valores de sus respuestas ya que era el mismo resultado de misma forma, se tenía escrito "Sí", "Si" y "Sí" en la que las tres significaban lo mismo pero lo trataban como una variable diferente, entonces lo que se hizo fue igualarlas todas a un mismo valor para evitar complicaciones de interpretación en el futuro. También, debido a que para nuestro análisis es crucial que la persona tenga marcada como "Si" la columna de Evaluado, lo que significa que tomó las pruebas, se eliminaron todas las filas donde esta columna tuviera como valor un NA, si bien se perdió bastante información de la base, alrededor de un 80%, esto el equipo lo considero como algo necesario ya que si no tenemos los resultados de esas pruebas no podemos interpretar resultados, de tal manera que si bien nos quedamos con pocos datos, al menos son datos que nos proporcionan la información que necesitamos.

Pasando a las columnas de las 5 pruebas pym metrics lo que se hizo fue cambiar sus resultados de evaluación por valores numéricos. "Do Not Recommend" = 1, "Recommend" = 3 y "Highly Recommend" = 5. Esto se hizo porque consideramos que es más fácil tratar con valores numéricos con significado categorico para poder interpretar nuestros datos.

En cuanto a las columnas de Apto/No Apto y Destacado de los resultados de las pruebas pym metrics lo que hicimos fue cambiar sus valores por 1 y 0 para hacerlo binario, en la columna de “Apto/No Apto” se le asignó a “Apto” = 1 y a “No Apto” = 0. Esto se hizo de manera similar para la columna “Destacado Pym” donde a “Destacado” = 1, y los valores vacíos o NaN se les asignó 0, esto no será de utilidad a la hora de querer utilizar modelos de análisis de datos.

Para la columna de Apto y Destacado de las pruebas del idioma inglés lo que se hizo fue juntar ambas columnas dentro de una sola y después asignarles un valor numérico, “Apto” = 1 y “Destacado” = 2.

En cuanto a los resultados de los niveles de inglés primero lidiamos con las variables de “False beginner” y “Late hangup” en la que dicha solución fue cambiar su valor a 0. Posteriormente se asignaron valores nuevos a los niveles de inglés de la siguiente manera.

- A1 = 1
- A2 = 2
- B1 = 3
- B2 = 4
- C1 = 5
- C2 = 6

Por último la variable de “Ingresados” decidimos cambiar sus valores por numéricos binarios, “Sí” = 1 y los valores vacíos se les asignó 0, esto una vez más nos servirá para poder realizar análisis de datos.

Aplicación de técnicas de modelación

Variables predictoras y variable target

Existen una gran variedad de modelos de clasificación que nos podrían ser de utilidad para tratar de encontrar una solución a nuestra problemática, pero independientemente de cual modelo sea el que nos resulte el más efectivo es importante primero el definir con qué valores estaremos trabajando para la construcción de los modelos. Para esto escogimos una serie de columnas que utilizaremos como nuestras variables predictoras y una variable de respuesta, dicho datos se pueden ver resumidos en la siguiente tabla.

Columna	Descripción	Tipo de dato	Predictora o Respuesta
Operaciones-Calidad	Prueba Pymmetric	Entero categórico	Predictor
MTTO-DIMA	Prueba Pymmetric	Entero categórico	Predictor
Comercial-Planeamiento	Prueba Pymmetric	Entero categórico	Predictor
DIGI-SC	Prueba Pymmetric	Entero categórico	Predictor
Resto-Soft	Prueba de “soft skills”	Entero categórico	Predictor
Destacado Pym	Indicador si el sujeto de prueba es destacado o no	Binario (0 ó 1)	Respuesta

Explicación de modelos de aprendizaje e hiperparametros

Los algoritmos de aprendizaje automático se enfocan en crear una función “f” para con los datos de entrada “x” poder predecir una variable dependiente “y”. Cada algoritmo funciona de diferente manera pero la premisa es la misma. Estos algoritmos tienen que ser entrenados y aprender, por lo general con grandes cantidades de datos. El tipo de aprendizaje de estos algoritmos puede ser supervisado (cuando se le da la variable “y”),

no supervisado (cuando no se le da la variable “y”) y aprendizaje por refuerzo. Por ejemplo, el algoritmo “K-Nearest Neighbors” se enfoca en graficar todos los datos que se le pasen en algo parecido a un gráfico de dispersión, cuando recibe los datos de los que se le quiera hacer predicciones, este los coloca en el gráfico de dispersión y revisa cuales son los vecinos más cercanos y hace sus predicciones en base a esto.

Los modelos de árboles de decisiones por clasificación se pueden enfocar en la medida de “entropía” para crear un diagrama de árboles con el cual se categoriza los datos por esta medida. En otras palabras, matemáticamente se genera un árbol de decisiones basado en la pérdida y la ganancia de información, una vez que se tiene el árbol, se puede seguir su camino a través de los nodos para poder hacer predicciones. Estos son solo 2 ejemplos, pero hay bastantes algoritmos de aprendizaje, cada uno con técnicas diferentes, sin embargo con la misma idea de aprender de una cantidad grande de información para poder hacer predicciones.

Cada uno de estos algoritmos tienen algo llamado “hiperparametros” estos parámetros son únicos y diferentes dependiendo del algoritmo. Son pequeñas medidas que le dicen al algoritmo como comportarse, y en base a ellos puede mejorar o empeorar la precisión del algoritmo. Por ejemplo, en el modelo mencionado anteriormente, “K-Nearest Neighbors”, uno de sus hiperparametros es el número de vecinos con el que se basa para hacer su predicción. Si le indicamos que queremos que haga su predicción en base a 5 vecinos, tomará en cuenta los 5 datos o nodos más cercanos para tomar su decisión predictiva, si le decimos que 10 vecinos, tomará en cuenta 10. Podemos jugar con esta métrica para ver cómo cambian las predicciones del algoritmo. Otros algoritmos como el árbol de decisiones tiene más y más complicados los hiperparametros, por ejemplo este tiene dos opciones de “gini” y de “entropía” que es la métrica que usará para hacer el árbol, tiene “profundidad del árbol”, tiene “máximo número de nodos”, etc. En conclusión se debe de tener un buen conocimiento del algoritmo y de sus hiperparametros para poder usarlo de la manera más efectiva y óptima.

Metodología de entrenamiento y prueba

Con las variables ya definidas podemos empezar a separar nuestro dataset en uno de entrenamiento y en uno de prueba, el punto de hacer esto es que el dataset de entrenamiento se encarga de aprender a predecir nuestra variable objetivo y una vez que el entrenamiento está completado se utiliza el dataset de prueba para verificar la veracidad de nuestras predicciones.

A la hora de completar nuestra limpieza de datos nos quedamos únicamente con 1050 aplicantes de los 5742 que había inicialmente, esto quiere decir que se perdió más del 80% de la información original, recalcar esto es importante ya que usualmente los parámetros más comunes para dividir los datasets es 80% del dataset original para entrenamiento y 20% de prueba o 70% de entrenamiento y 30% de prueba, pero dado que nuestra base se vió muy resumida utilizaremos de una distribución de 80%-20% para entrenamiento y prueba de forma respectiva.

Métricas de evaluación

Ya con las variables definidas y los parámetros de entrenamiento y prueba definidos, cada integrante probó diferentes modelos en los cuales se obtuvieron diferentes resultados, pero antes de eso para evitar confusión hay que explicar algunos conceptos que aparecen en la tabla.

El primer valor a mencionar es el de “Precisión”, el modelo solo puede predecir dos tipos de valores, verdadero o falso, que cuando lo comparamos a nuestra situación lo que trata de predecir si es destacado o no. Entonces cuando hablamos de precisión nos referimos al porcentaje de veces que nuestro modelo acierta cuando predice un valor como verdadero.

Para falsos positivos nos referimos al porcentaje de veces que nuestro modelo se equivoca y predice un valor como destacado cuando en realidad no lo era. Luego en sensitividad es el porcentaje es su capacidad de predecir un valor de tipo destacado cuando efectivamente era alguien destacado. Para el error de clasificación estamos hablando en términos generales que tanto se equivoca el método. Para especificidad se

refiere a la capacidad de predecir correctamente a personas no destacadas. En cuanto a F1 es una métrica que generaliza a la precisión y la especificidad.

La curva ROC, es una forma de representar gráficamente la habilidad de un clasificador binario. Demuestra la relación entre sensibilidad y especificidad. Entre más recta esté la curva (γ) mejor. El indicador Roc Auc, es una medida del área bajo la curva ROC, así es más fácil comparar gráficas similares, entre más cerca esté de 1, mejor.

También es necesario hacer algún tipo de validación cruzada dado que las pruebas anteriores son solamente usando los datos de prueba pero es necesario checar cómo se comportará el modelo cuando ya no se traten estos datos sino se esté manejando una base de datos independiente. Existen una gran variedad de técnicas para lograr esto pero la que se escogió en esta ocasión fue K-fold validation en el cual lo que se hace es dividir nuestros datos en una serie de k-subconjuntos donde uno de los subconjuntos se utiliza como prueba y el resto se usa como entrenamiento, este proceso se repite k veces cambiando el subconjunto que se usa como prueba, para cada iteración se obtiene la precisión de cada uno y al final se suman y se obtiene su promedio.

Generación de modelos y evaluación en base a las diferentes métricas

Para los modelos que se podrán observar en la tabla de resultados, estos fueron aquellos que para cada integrante del equipo les proporcionó los mejores resultados, lo que se hizo después de seleccionar estos 3 modelos de mejor desempeño pasamos a realizar un ajuste de hiperparametros para ver si era posible él mejorar un tanto más la predicción de nuestros modelos.

El primero es el de Bosques aleatorios en el cual los hiperparametros utilizados fueron el de “n_estimators” el indica el número de árboles que habrá en el bosque, luego es el de “max_features” que nos indica el número máximo de características que tenga un nodo de división, luego está “max_depth” el cual determina el número de decisiones máximas por cada árbol y por último tenemos el “criterion” que nos indica que tan bueno es el criterio para crear la partición de decisión.

Para el modelo de árboles de decisión comparte los mismos hiperparámetros que bosques aleatorios en “max_depth”, “max_features” y “criterion”, el único que es diferente en esta ocasión es “min_samples_leaf” que nos indica el número de nodos mínimos requerido para tener un nodo hoja.

Pasando al método de k-vecinos más cercanos primero tenemos el “n_neighbors” que nos indica el número de vecinos a intentar buscar, luego está “algorithm” que es la forma en la que buscará los vecinos, posteriormente está “leaf_size” el cual afecta la construcción y memoria utilizada para hacer el árbol, después está “metric” qué es lo que determina si se utiliza una distancia de manhattan, minkowski o euclídea, y por último se tiene “p” que determina el valor que tomará minkowski, si $p = 1$ entonces utiliza la distancia manhattan y si $p = 2$ entonces utiliza euclídea, dado que tenemos que agregar el valor de “p” podemos de aquí saber que se está utilizando minkowski.

Ya con estos hiperparámetros definidos podemos mostrar los resultados que nos dieron nuestros intentos individuales.

Integrante	Modelo 1 Resultados	Modelo 2 Resultados	Modelo 3 Resultados
Andres Saldaña	Bosques Aleatorios Exactitud: 89% Precisión: 87% Sensitividad: 90% Especificidad: 88% Roc Auc: 96% K-Fold: 85% +/-15%	Árboles de Decisión Exactitud: 88% Precisión: 88% Sensitividad: 87% Especificidad: 89% Roc Auc: 93% K-Fold: 83% +/-17%	K-Vecinos Exactitud: 85% Precisión: 83% Sensitividad: 86% Especificidad: 84% Roc Auc: 92% K-Fold: 84% +/-15%
Eugenio Morales	Bosques Aleatorios Precisión: 85% Falsos positivos: 18% Sensitividad: 90% K-Fold: 85%(+27%) Error Clasificación: 13% Especificidad: 82% F1: 88% AUC: 95%	Árboles de Decisión Precisión: 83% Falsos positivos: 20% Sensitividad: 87% K-Fold: 82%(+27%) Error Clasificación: 16% Especificidad: 79% F1: 85% AUC: 92%	K-Vecinos Precisión: 85% Falsos positivos: 15% Sensitividad: 83% K-Fold: 78%(+17%) Error Clasificación: 21% Especificidad: 84% F1: 79% AUC: 92%

Karen González	Bosques Aleatorios Exactitud: 80% Precisión: 57.8% Sensitividad: 53.1% Especificidad: 88.2% Roc Auc: 89.9%	Árboles de Decisión Exactitud: 80% Precisión: 59% Sensitividad: 46.9% Especificidad: 90.1% Roc Auc: 80.6%	Gradient Booster Exactitud: 77.6% Precisión: 52.5% Sensitividad: 42.9% Especificidad: 88.2% Roc Auc: 86.6%
Fernando González	Bosques Aleatorios Exactitud: 83% Precisión: 62% Sensitividad: 49% Especificidad: 87% Roc Auc: 88%	Árboles de Decisión Exactitud: 79% Precisión: 58% Sensitividad: 53% Especificidad: 89% Roc Auc: 82%	K-Vecinos Exactitud: 84% Precisión: 57% Sensitividad: 47% Especificidad: 88% Roc Auc: 89%
Daniel Sánchez	Bosques Aleatorios Exactitud: Precisión: Sensitividad: F1: ROC:	Árboles de Decisión Exactitud: Precisión: Sensitividad: F1: ROC:	Regresión Logística Exactitud: Precisión: Sensitividad: F1: ROC:

Viendo estos resultados y observando las otras variables de nuestro dataset, notamos que teníamos un sesgo en la categoría “Apto/No Apto” la cual causaba que los modelos desempeñarán demasiado bien como para que esto se considerara factible, entonces, para poder hacer una comparación agregamos esta variable a nuestro modelo y los resultados fueron los siguientes.

Integrante	Modelo 1 Resultados	Modelo 2 Resultados	Modelo 3 Resultados
Andres Saldaña	Bosques Aleatorios Exactitud: 99% Precisión: 99% Sensitividad: 100% Especificidad: 99% Roc Auc: 99% K-Fold: 97%(+3%)	Árboles de Decisión Exactitud: 99% Precisión: 99% Sensitividad: 100% Especificidad: 99% Roc Auc: 99% K-Fold: 98%(-2%)	K-Vecinos Exactitud: 99% Precisión: 98% Sensitividad: 99% Especificidad: 98% Roc Auc: 99% K-Fold: 91%(-9%)
Eugenio Morales	Bosques Aleatorios Precisión: 98% Falsos positivos: 1% Sensitividad: 100 % K-Fold: 99%(+2%) Error Clasificación:	Árboles de Decisión Precisión: 97% Falsos positivos: 2% Sensitividad: 96 % K-Fold: 97%(+5%) Error Clasificación: 3%	K-Vecinos Precisión: 98% Falsos positivos: 1.2% Sensitividad: 98 % K-Fold: 93%(+13%) Error Clasificación: 1%

	0.6% Especificidad: 98% F1: 99% AUC: 99%	Especificidad: 97% F1: 97% AUC: 99%	Especificidad: 98% F1: 98% AUC: 99%
Karen González	Bosques Aleatorios Exactitud: 98.6% Precisión: 94% Sensitividad: 100% Especificidad: 98.2% Roc Auc: 98.9%	Árboles de Decisión Exactitud: 98.6% Precisión: 94% Sensitividad: 100% Especificidad: 98.2% Roc Auc: 98.9%	Gradient Booster Exactitud: 98.6% Precisión: 94% Sensitividad: 100% Especificidad: 98.2% Roc Auc: 98.6%
Fernando González	Bosques Aleatorios Exactitud: 99.09% Falsos positivos: 9% Sensitividad: 100% Especificidad: 98.27% F1: 99.05% Roc Auc: 99.11%	Árboles de Decisión Exactitud: 99.09% Falsos positivos: 0.9% Sensitividad: 100% Especificidad: 98.27% F1: 99.05% Roc Auc: 99.11%	K-Vecinos Exactitud: 93.95% Falsos positivos: 6% Sensitividad: 97.45% Especificidad: 90.80% F1: 93.86% Roc Auc: 96.77%
Daniel Sánchez	Bosques Aleatorios Exactitud: 99% Precisión: 99% Sensitividad: 100% F1: 99% ROC: 100%	Árboles de Decisión Exactitud: 99% Precisión: 98% Sensitividad: 97% F1: 98% ROC: 97.9%	Regresión Logística Exactitud: 98.6% Precisión: 97% Sensitividad: 98% F1: 98% ROC: 99.7%

Cómo se puede observar, estos resultados son buenos, demasiado buenos al punto que uno puede considerar que no son verdad, el motivo de estos resultados pueden ser varios factores tales como que no haya suficiente información de datos para que el modelo pueda trabajar de buena manera, la explicación y teoría principal que tenemos es que dicha variable está muy sesgada con relación a nuestro modelo lo cual causa que estos datos tengan valores tan altos a los que uno esperaría, esta comparación sirve de ejemplo de las consecuencias que puede tener el trabajar con datos sesgados.

Evaluación

Evaluación de Resultados

Tras haber probado los diferentes algoritmos de manera independiente, observamos que obtuvimos resultados muy similares en nuestros modelos de mayor desempeño, siendo

estos K-Vecinos, Bosques Aleatorios, y Árboles de decisión. Tomando en cuenta que en cuestiones de métricas, los algoritmos mencionados se desempeñaron de una forma muy similar y debido a nuestro conocimiento de los modelos es que decidimos utilizar “Bosques aleatorios” como nuestro modelo definitivo dado que es el que sentimos que podemos trabajar de mejor manera.

Es importante mencionar que durante el proceso de construcción del modelo, los datos fueron escalados o normalizados. Esto quiere decir que todos los datos tenían un rango muy parecido. Ya que como había columnas con datos más grandes, estos iban a tomar un peso mayor, sobre las columnas con datos más pequeños. Sin embargo, durante el desarrollo de la aplicación web del modelo, vimos que no teníamos porque escalar los datos. Esto porque cuando un usuario ingresa los datos, estos mismos no tenían el mismo escalamiento de los datos con los que entrenamos el modelo, lo que ocasionó que esta predicción fuera errónea. Tomando en cuenta que cuando entrenamos al modelo lo entrenamos con 0.25, 0.35, por poner un ejemplo, y cuando quisimos hacer las predicciones le estábamos dando 1 y 0. Al final, cuando se decidió remover esta parte del escalamiento de los datos del programa, vimos que tuvo un efecto nulo. Es decir, pasamos de un 98% de precisión a un 97.5% por poner un ejemplo, entonces dado el ínfimo cambio que esto hacía podemos decir que no hay un impacto fuerte en el remover el escalamiento.

Impacto social

El impacto social que podemos alcanzar con la realización de nuestra solución se podrá ver reflejada en el bienestar de los trabajadores de Ternium. La razón por la que consideramos por la cual este es el caso es debido a que si no realiza la buena selección de nuevo personal de trabajo se pueden generar dos escenarios principales.

El primer caso puede ser que la persona seleccionada para dicho puesto a final de cuentas esté por debajo del nivel solicitado para la sección a la cual fue asignado, esto lo que puede llegar a causar es ansiedad sobre el trabajador debido a que siente que no es lo suficiente para poder desempeñar de manera correcta en su trabajo, posteriormente esa ansiedad puede convertirse en estrés lo cual puede ocasionar un

daño aún mayor en la salud mental del empleado lo cual podría tener consecuencias en su trabajo y entonces se crea un ciclo donde principalmente el empleado se ve afectado.

El otro sería en el polo opuesto donde la persona seleccionada para el puesto terminó estando sobrecalificada para la zona en la cual fue asignado, de manera similar al caso anterior el estar realizando un trabajo el cual es muy poco para la persona puede hacerla sentir encerrada, generar emociones de frustración, estrés, sentirse incompleto, todo esto de igual manera la que representa es un daño en la salud mental del trabajador.

Sabiendo esto podemos entonces reiterar que nuestra solución que busca la selección calificada del personal adecuado para cada puesto puede tener un impacto social, en este caso dicho impacto sería que el poder asegurarnos de que se está contratando a la persona correcta, podemos cuidar de mejor manera la salud mental de los empleados ya que se está teniendo una mejor seguridad de que están laborando en un lugar donde pueden desempeñar de mejor manera.

Objetivos de desarrollo sostenible

A la hora de la realización de un proyecto es siempre importante considerar si nuestras acciones son capaces de cumplir hacia los objetivos de desarrollo sostenible, que si bien dependiendo del proyecto en cuestión esto puede ser el caso en mayor o menor medida, siempre es importante tenerlo en consideración. Siendo este el caso podemos ver que tipo de impacto nuestra solución puede tener.



Dado que nuestro proyecto se centra en el área de recursos humanos en la tarea de reclutamiento de nuevo personal, es bastante probable que exista sesgo en los criterios de evaluación y de selección de nuevo personal, dada la situación que se vive hoy en día es indudable que dicho sesgo en cierto porcentaje sea por cuestiones de género, de esta forma el desarrollar un método de selección de personal el cual se encuentre menos sesgado nos ayudará a reducir estas desigualdades de trabajo por cuestiones de género.

8 TRABAJO DECENTE Y CRECIMIENTO ECONÓMICO



La selección incorrecta de personal para una área puede resultar dañina tanto para el bienestar del trabajador como para el rendimiento y desempeño de la compañía, entonces el poder mejorar el proceso de selección de empleados a aquellos que se encuentren correctamente calificados nos asegurara que la persona tenga una línea de trabajo decente acorde a sus habilidades, entonces el tener un equipo de trabajo que realiza su actividades de manera correcta mejora el rendimiento de la empresa lo cual tiene un impacto en el crecimiento económico.

10 REDUCCIÓN DE LAS DESIGUALDADES



Como bien se mencionó en la ODS 5, existe sesgo a la hora de contratación por cuestiones de género, pero la realidad es que este sesgo no es el único existente, también pueden existir otros factores como tipo de educación recibida, instituciones de educación donde el aplicante estudio, así como otras cuestiones sociales que se pueden estar tomando en cuenta ya sea de manera intencional o no intencional. Es por eso que crear un mejoramiento al proceso de selección de personal donde únicamente las variables importantes son consideradas ayudará a generar una fuerte reducción en desigualdades de trabajo.

Si bien no parecen ser muchos objetivos a cumplir en esta ocasión, recordar que debido a la naturaleza específica del proyecto en cuestión era de esperarse que su ayuda no pudiera ser tanta en comparación con otros proyectos. Aún siendo este el caso, al final de todo esto se está generando un impacto en estos objetivos, lo que es lo que importa realmente.

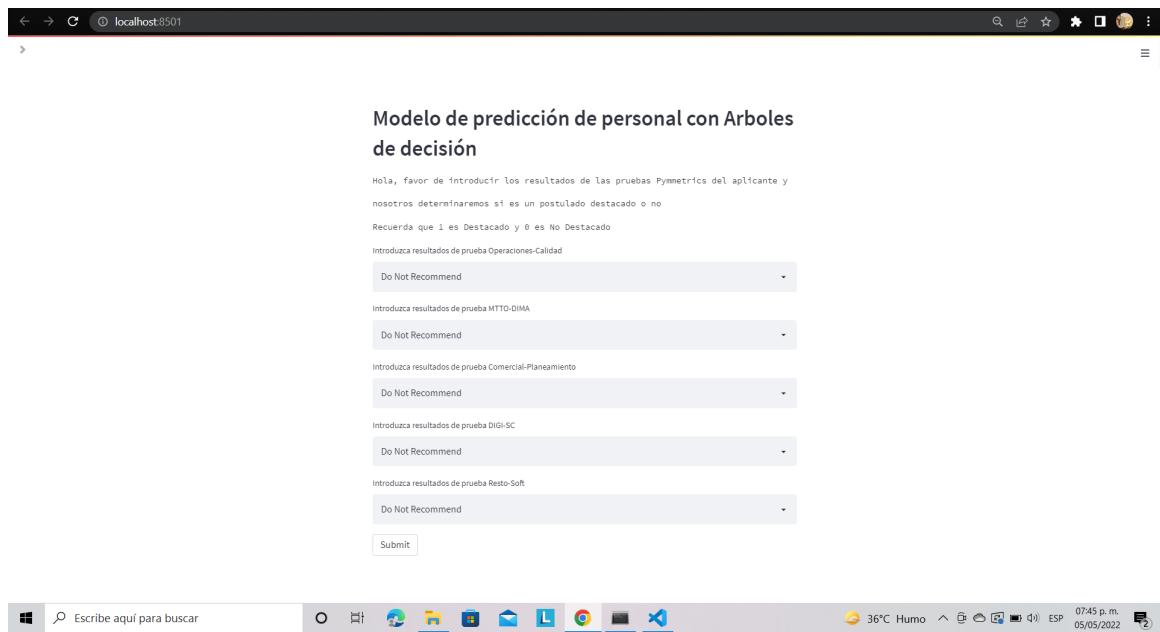
Despliegue

Descripción del prototipo principal

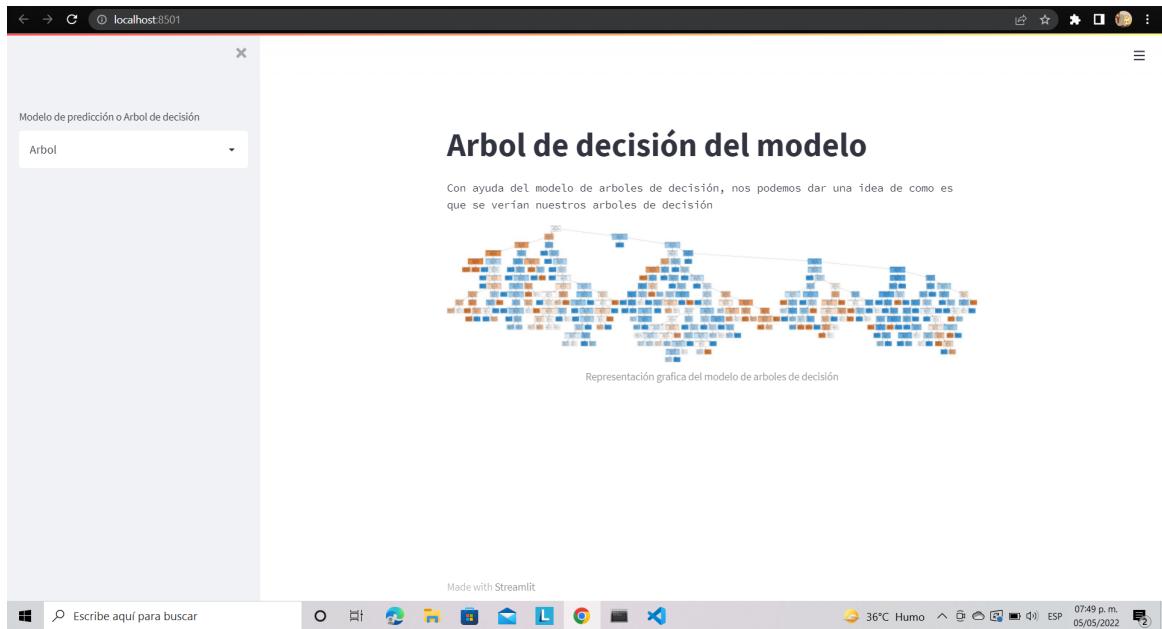
Como prototipo inicial de nuestra app de desarrollo decidimos generar una aplicación en línea con la ayuda de Streamlit el cual es un sitio de uso libre para la generación de sitios web para modelos de machine learning para que estos sean más interactivos.

De momento la idea es simple, lo que se hizo fue generar utilizar nuestro modelo de selección previamente escogido y desarrollar un código que nos ayude a desplegarlo de una manera más visualmente atractiva, esto fue con ayuda de un tutorial encontrado en internet [8].

En sí lo que se despliega en pantalla es una serie de botones con los resultados de las pruebas realizadas que en sí son nuestras variables que utilizamos para crear nuestro modelo de predicción, entonces lo que uno hace es seleccionar el resultado que obtuvo el aplicante en cuestión, presionar el botón de respuesta y el programa lo que nos devuelve es un 1 ó 0, donde uno se refiere a que es Destacado y 0 que no lo es. Esta es una vista preliminar de cómo se ve la página principal.



Como bien mencionamos, se encuentra en etapas iniciales y es por eso que de momento la forma se encuentra de manera muy sencilla, también como pequeño bonus pusimos un submenú que al ser ingresado nos puede mostrar una representación gráfica de cómo es que se vería nuestro modelo si se utilizan árboles de decisión.



Recomendaciones

Recomendaciones al negocio

Una recomendación que podemos mencionar es que con este entregable junto con los entregables del resto de los equipos hayan servido de ayuda para darse cuenta que el uso de la ciencia de datos es algo que crece de manera exponencial conforme pasa el tiempo una gran cantidad de empresas ya se encuentra iniciando en este campo laboral, uno en sitios de redes de trabajo como los es Linkedin se puede observar una vasta cantidad de compañías buscando gente con conocimientos en ciencia de datos debido a la alta demanda que esto tiene. Entonces a forma de recomendación sugerimos que Ternium comience a investigar sobre esta área, y, de forma gradual a invertir en un

departamento el cual se pueda dedicar a la ciencia de datos, pues, después de todo ya pudieron observar cómo esta área puede resultar de mucha utilidad para fortalecer las áreas de oportunidad cuya solución recae en la ciencia de datos y así colocar a Ternium como una industria 4.0.

Recomendaciones técnicas

Si bien consideramos que la estructura de la base de datos es acertada a la hora de cuantificar las distintas métricas y pruebas mediante las cuales los se valoran las competencias de los postulantes, encontramos que la falta de un diccionario para las categorías de la base de datos entorpeció en diversas ocasiones nuestro análisis de la misma, pues, esto dificulta la diferenciación entre aquellas categorías cuyo fin es únicamente identificar a la persona de las que nos son útiles para un análisis predictivo.

También, se propone el establecimiento de un estándar para el ingreso de datos para las bases de datos para así evitar tener valores los cuales representan la misma características, pero, que por estar escrito ligeramente distinto, estos no puedan ser reconocidos como tal. Es decir, establecer que si se tiene un valor categórico con “si/no” estos no sean escritos algunos con acentuación y otros no, o, con un espacio al final (“si ”).

Siguientes pasos

En cuanto a siguientes pasos a realizar, como se habrán dado cuenta la naturaleza de estos proyectos el tiempo es algo esencial para poder desarrollar un resultado que sea satisfactorio tanto para nosotros como para ustedes, como pueden ver el resultado de trabajo en estas 5 semanas se puede ver reflejado en el documento y la presentación por parte de nuestros equipo, pero de igual manera podemos decir con certeza que nuestro producto podría ser mejor a lo que es ahora y para eso dos cosas son esenciales, tiempo y más datos. Usualmente el 60% del tiempo de los proyectos de ciencia de datos suele ser destinado a tan solo la limpieza de datos, lo cual en nuestro contexto de 5 semanas no nos da mucho espacio para trabajar en el modelaje,

evaluación, ajuste de parámetros e implementación en una aplicación web, si bien logramos dar un resultado que nosotros consideramos de calidad, este se encuentra en sus primeras etapas y podría ser mejor.

El segundo aspecto a mencionar es el de los datos, para poder nosotros mejorar nuestro modelo de predicción requerimos de más datos por parte de la empresa para poder funcionar, en el transcurso de este documento se mencionaron conceptos como sesgo, falta de datos, pérdida de datos, desbalance de datos, esto se menciona por dos factores principales, la cantidad y calidad de los datos, la gran mayoría de la base de datos del documento original eran campos vacíos o incompletos o con opciones de llenado que no eran acorde al campo que se estaba tratando, es por eso que a la hora de tratar la limpieza de datos tuvimos que eliminar gran parte de la información. Somos conscientes que no fue hasta que empezaron a colaborar con nosotros que se dieron cuenta que la forma en la que registraban a sus aspirantes no era la óptima y esta puede estar de una manera más organizada así que realmente bajo ninguna circunstancia los estamos culpando del desempeño de nuestro trabajo, sino lo que queremos es hacerles saber para futuros trabajos que es imprescindible el contar con una base de datos mejor estructurada para que así las personas encargadas de hacer los modelos no tengan tantas complicaciones.

Una segunda acción que se puede realizar es que ya se vió como la ciencia de datos puede ayudar al reclutamiento de personal, entonces puede surgir la idea de utilizarlo en otros ámbitos que inclusive puede seguir siendo de recursos humanos. Ya se habló de reclutamiento entonces también se puede hablar de posibles renuncias, es decir generar un modelo que pueda predecir si un trabajador va a renunciar o no y en base a eso ustedes decidir cómo manejar la situación para poder mantener a sus empleados. Fuera de recursos humanos esto también puede ser utilizado para predicciones de ventas, ganancias, producción, etc. Si bien esto entra ahora en algo conocido como modelos de regresión, también es posible utilizar la ciencia de datos para predecir este tipo de ámbitos.

Entonces como pueden ver las opciones son vastas y todo depende de que es en lo que se quieren enfocar, lo único que hay que mencionar es que para cada modelo de machine learning que quieran utilizar, va a ser necesario hacer una investigación

exhaustiva de qué tipo de datos son los que se tendrían que recolectar y qué importancia tiene en las consecuencias del mundo real y de ahí generar un dataset primerizo de los datos reunidos. Posteriormente a eso se tendría que realizar todo el proceso de igual manera al que se realizó para la modelación del caso de este documento rector.

En cuanto a siguientes pasos a realizar por parte de Ternium pueden ser varios, uno de ellos puede ser el uso de otras formas para poder registrar los datos de sus aplicantes, ya que como se vió y platicó durante el desarrollo de este proyecto la mala captura de los datos, una opción puede ser crear algún tipo de sitio en línea donde los aplicantes registren sus datos junto con los resultados de las pruebas y de ahí estos sean guardados de forma ordenada y automática en un base de datos para evitar errores en el registro de información, de ahí ya sería más fácil tener una accesibilidad a los datos y facilitaría la implementación del uso del modelo. Se que esto es más fácil decirlo que hacerlo y una disculpa si no podemos dar una respuesta más concreta de momento ya que eso implicaría tiempo para investigar y crear una idea más concreta de una solución más tangible, simplemente mencionamos esta idea como punto de partida para que a su consideración vean si deciden el mejorar dicha idea o buscar algún otro tipo de solución.

Fuentes bibliográficas

- [1] Huda, A., & Ardi, N. (2021). Predictive Analytic on Human Resource Department Data Based on Uncertain Numeric Features Classification. *International Journal of Interactive Mobile Technologies (iJIM)*, 15(08), pp. 172–181.
<https://doi.org/10.3991/ijim.v15i08.20907>
- [2] Vulpen, E. V. (13 de octubre de 2013). *What is HR analytics? Human Resources Analytics [updated 2021]*. What is HR Analytics? Retrieved 2 de abril de 2022 de
<https://www.aihr.com/blog/what-is-hr-analytics/>
- [3] Portafolio. (2021, 27 abril). *Reclutamiento de personal se apalanca en la ciencia de datos*. Recuperado 2 de abril de 2022, de
<https://www.portafolio.co/economia/empleo/reclutamiento-de-personal-se-apalanca-en-la-ciencia-de-datos-551370>
- [4] Belizón, M. J., & Kieran, S. (2021). Human resources analytics: A legitimacy process. *Human Resource Management Journal*.
- [5] Hampel, C., Lawrence, T., & Tracey, P. (2017). Institutional work: Taking stock and making it matter. In R. Greenwood, C. Oliver, & T. B. Lawrence (Eds.), *The SAGE handbook of organizational institutionalism* (pp. 558–590). Sage
- [6] Bao, L. (2021). *Big data analytics en la Gestión de Recursos Humanos*. LinkedIn. Recuperado de
https://www.linkedin.com/pulse/big-data-analytics-en-la-gesti%C3%B3n-de-recursos-humanos-liliana-bao?trk=public_profile_article_view#:~:text=La%20ciencia%20de%20datos%20es,la%20satisfacci%C3%B3n%20de%20los%20empleados.
- [7] INEGI, I. N. E. G. I. (2021, June 25). *Estadística a propósito del Día Mundial del ... - INEGI. ESTADÍSTICAS A PROPÓSITO DEL DÍA DE LAS MICRO, PEQUEÑAS Y MEDIANAS EMPRESAS (27 DE JUNIO) DATOS NACIONALES*. Recuperado de
<https://www.inegi.org.mx/contenidos/saladeprensa/aproposito/2020/MYPIMES20.pdf>

[8] Christian Ongko, G. (Febrero 18). Building a Machine Learning Web Application using Streamlit. Artículo en línea. Extraído de <https://towardsdatascience.com/building-a-machine-learning-web-application-using-streamlit-8c3d942f7b35>